# Maximum Likelihood Estimation of Population Parameters

**Yun-Xin Fu and Wen-Hsiung Li**

*Center for Demographic and Population Genetics, University of Texas, P.O. Box 20344, Houston, Texas 77225*

Manuscript received November 3, 1992
Accepted for publication April 5, 1993

ABSTRACT

One of the most important parameters in population genetics is $\theta = 4N_e\mu$ where $N_e$ is the effective population size and $\mu$ is the rate of mutation per gene per generation. We study two related problems, using the maximum likelihood method and the theory of coalescence. One problem is the potential improvement of accuracy in estimating the parameter $\theta$ over existing methods and the other is the estimation of parameter $\lambda$ which is the ratio of two $\theta$'s. The minimum variances of estimates of the parameter $\theta$ are derived under two idealized situations. These minimum variances serve as the lower bounds of the variances of all possible estimates of $\theta$ in practice. We then show that Watterson's estimate of $\theta$ based on the number of segregating sites is asymptotically an optimal estimate of $\theta$. However, for a finite sample of sequences, substantial improvement over Watterson's estimate is possible when $\theta$ is large. The maximum likelihood estimate of $\lambda = \theta_1/\theta_2$ is obtained and the properties of the estimate are discussed.

CONSIDER a gene (locus) in a random mating population of effective population size $N_e$. Let $\mu$ be the mutation rate per gene per generation at the locus. As is well known, the parameter $\theta = 4N_e\mu$ plays a prominent role in the stochastic theory of population genetics. Therefore, accurate estimation of this quantity is important. A closely related problem is the estimation of the ratio, $\lambda$, of two $\theta$'s. In the case that the effective sizes of the two populations are the same, $\lambda$ is the ratio of mutation rates. A special case is the ratio of the rate of non-synonymous substitution and the rate of synonymous substitution in a given gene.

Using the infinite site model (KIMURA 1969), WATTERSON (1975) derived the distribution of the number of segregating sites, $K$, in a random sample of $n$ genes from a single random mating population. The expectation and the variance of $K$ are

$$E(K) = \theta a_n \tag{1}$$

$$\mathrm{Var}(K) = \theta a_n + \theta^2 b_n \tag{2}$$

where

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i} \quad \text{and} \quad b_n = \sum_{i=1}^{n-1} \frac{1}{i^2}. \tag{3}$$

TAJIMA (1983) considered the number ($\prod$) of nucleotide differences between a random pair of genes. He showed that the average, $\prod_n$, of $n(n-1)/2$ pairwise $\prod$ values has the following expectation and variance:

$$E(\prod_n) = \theta \tag{4}$$

$$\mathrm{Var}(\prod_n) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2. \tag{5}$$

The two statistics $K/a_n$ and $\prod_n$ can be used to estimate the value of $\theta$. When $\mu$ is known, it is equivalent to estimating the effective population size; and when $N_e$ is known, it is equivalent to estimating the mutation rate $\mu$.

Because the variance of $K/a_n$ is smaller than the variance of $\prod_n$, $K/a_n$ (WATTERSON's estimator) is a better estimator of $\theta$ than $\prod_n$. However, despite the importance of the issue, little improvement has been made over Watterson's estimator. In fact, it is not even clear whether substantial improvement in the accuracy of estimation of $\theta$ is possible. The improvement is measured in terms of the variance of the estimator of $\theta$. Under the assumptions that sequences are infinitely long and that the scaled coalescent times can be estimated without error, FELSENSTEIN (1992) showed that the improvement in accuracy is not only possible but can be to such an extent that the efficiency of WATTERSON's estimator (i.e., the ratio of the variance of FELSENSTEIN's hypothetic estimator and the variance of $K/a_n$) approaches zero as the sample size becomes very large.

Although FELSENSTEIN's (1992) assumptions are not realistic, his study raises the question of how much improvement on the estimation of $\theta$ can be made in practice. This is a difficult question, and so an alternative approach is to know the largest lower bound of the variances of all possible estimators of $\theta$ that can be achieved in practice. This can serve as a measure of the closeness of a new estimator to optimality. This knowledge can tell us whether further efforts to improve the accuracy in estimation of $\theta$ are worthwhile. The main purpose of the paper is to provide by the maximum likelihood approach a more realistic lower

bound of the variances of estimators of $\theta$ than that of FELSENSTEIN (1992). Note that we are not proposing any new practical estimator of $\theta$ but are only addressing the issue of how much better one might be able to do compared to WATTERSON's estimator. Another purpose is to study the maximum likelihood estimate of the ratio $\lambda$ of two $\theta$'s.

## JOINT DENSITY FUNCTIONS OF EVOLUTIONARY EVENTS

Suppose a random sample of $n$ genes is taken from a single random mating population. Under the assumption of neutral mutations, the process governing the evolution of the sequences being sampled is entirely determined by the value of $\theta$. This process, which is often called the coalescence process, has been studied by KINGMAN (1982), HUDSON (1982) and TAJIMA (1983).

We use the Wright-Fisher model for the population and assume no recombination between sequences. Then the $n$-coalescent time, $t_n$, that a sample of $n$ genes were derived from $n-1$ distinct ancestors $t_n$ generations ago follows the exponential distribution.

$$g\left(t_n; \frac{4N_e}{n(n-1)}\right) = \frac{n(n-1)}{4N_e} \exp\left(-\frac{n(n-1)}{4N_e} t_n\right). \quad (6)$$

For each sample of $n$ genes, there is a genealogy which connects the $n$ genes to their single common ancestor. We shall assume that the number of mutations, $k$, in a gene for a given time interval of length $t$ follows the Poisson distribution:

$$p(k; \mu t) = \frac{\exp(-\mu t)(\mu t)^k}{k!}. \quad (7)$$

Consider the genealogy of a random sample of $n$ genes from a single random mating population (e.g., Figure 1). There are $n-1$ internal nodes in the tree numbered from 2 to $n$ according to their order of occurrence in time. Therefore, between the $(i-1)$th and the $i$th node there are exactly $i$ branch segments, labeled from 1 to $i$. The time from the $(i-1)$th node to the $i$th node is the coalescent time $t_i$. Define $\eta_{ij}$ as the number of mutations occurring in the $j$th branch of the $i$ branch segments between the $(i-1)$th node and the $i$th node.

It is reasonable to assume that the spacial distribution of a given number of mutations among the sites of a gene is independent of the parameter $\theta$, though the number of mutations is dependent on $\theta$. For example, one may assume that the spacial distribution is uniform among all the sites, or any distribution that does not contain the parameter $\theta$. Therefore, all relevant information about the value of $\theta$ in a sample of $n$ genes is contained in the vector

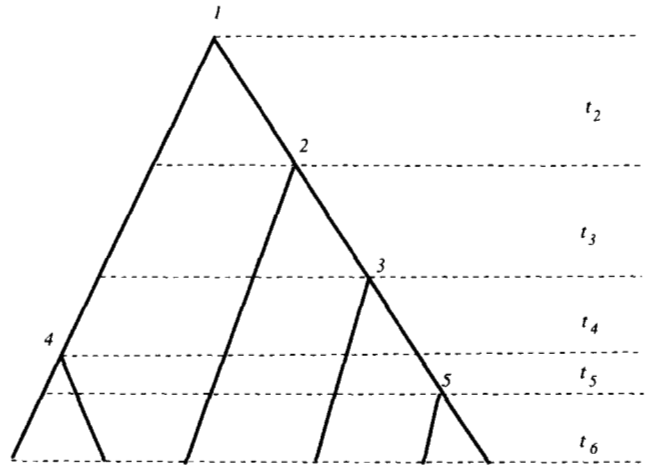


FIGURE 1.—An example of genealogy of a sample of six genes. The number beside each node is the branching order of the node and $t_n$, $i = 2, \ldots, 6$ are the number of generations between successive branching points.

$$\Theta = \{t_m, \eta_{mj}: j = 1, \ldots, m; m = 2, \ldots, n\}$$

which contains $(n-1)(n+4)/2$ elements. The joint probability density function (the likelihood function) of the quantities in $\Theta$ is

$$f(\Theta) = \prod_{m=2}^{n} \left(\prod_{i=1}^{m} p(\eta_{mi}; \mu t_m)\right) g\left(t_m; \frac{4N_e}{m(m-1)}\right) = \prod_{m=2}^{n} p(\eta_m; \mu m t_m) g\left(t_m; \frac{4N_e}{m(m-1)}\right) \quad (8)$$

where

$$\eta_m = \sum_{i=1}^{m} \eta_{mi} \quad (9)$$

is the total number of mutations that occurred between the $(m-1)$th node and the $m$th node, i.e., during the m-coalescent time $t_m$. From (8), one can see that the vector $\Phi = \{t_m, \eta_m; m = 2 \ldots, n\}$ is a sufficient statistic for the parameter $\theta$. That is, all relevant information about the value of $\theta$ is contained in these $2(n-1)$ quantities.

From (8), one can obtain the joint density function of $(\eta_2, \ldots, \eta_n)$ by integration over coalescent times. Since it is well known that a Poisson variable with a parameter that is exponentially distributed is equivalent to a geometrically distributed random variable, it can be shown that the joint density function of $(\eta_2, \ldots, \eta_n)$ is

$$f(\eta_2, \ldots, \eta_n) = \prod_{m=2}^{n} \left(\frac{1}{\beta_m + 1}\right)\left(\frac{\beta_m}{\beta_m + 1}\right)^{\eta_m} \quad (10)$$

where

$$\beta_m = \theta/(m-1). \quad (11)$$

From the density function (10), one can obtain by convolution the density function of the total number of mutations in the genealogy, which is given by

$$\eta = \sum_{m=2}^{n} \eta_m. \tag{12}$$

The expectation and variance of $\eta$ can be obtained from the moment generation function of $\eta$ as WATTERSON (1975) did. A simpler way to derive them is through conditional expectations as follows:

$$\begin{aligned} E(\eta) &= \sum_m E(\eta_m) = \sum_m E_{t_m}(E(\eta_m | t_m)) \\ &= \sum_m \mu m E(t_m) \\ &= \theta a_n \end{aligned} \tag{13}$$

$$\begin{aligned} \mathrm{Var}(\eta) &= E(\eta^2) - E^2(\eta) \\ &= E_{t_2,\ldots,t_n}(E(\eta^2 | t_2, \ldots, t_n)) - E^2(\eta) \\ &= E\left[\mu \sum_m (m-1)t_m + \mu^2 \left(\sum_m m t_m\right)^2\right] - E^2(\eta) \\ &= \theta a_n + \theta^2 b_n. \end{aligned} \tag{14}$$

Define

$$\hat{\theta}_w = \eta / a_n. \tag{15}$$

Then, $\hat{\theta}_w$ is an unbiased estimate of $\theta$ as can be seen from Equation 13 and the variance is, from Equation 14,

$$\mathrm{Var}(\hat{\theta}_w) = \frac{\theta}{a_n}\left(1 + \frac{\theta b_n}{a_n}\right). \tag{16}$$

Note that if the infinite site model is assumed, then $\hat{\theta}_w$ is equivalent to the estimator $K/a_n$, because in this situation the total number of mutations equals to the number of segregating sites in the sample of genes. For this reason, we shall call the estimator $\hat{\theta}_w$ WATTERSON's estimator of $\theta$.

## LOWER BOUNDS OF THE VARIANCES OF ESTIMATORS OF $\theta$

For subsequent discussions to be meaningful, let us agree that the best one can do is to be able to observe the whole process of evolution of the set of sequences from the common ancestral node to the present. In this idealized situation, one can count not only the number of mutations in each branch (and thus $\{\eta_m, m = 2, \ldots, n\}$) but also the number of generations $\{t_m, m = 2, \ldots, n\}$ between successive branching events. An optimal estimator of $\theta$ with these two sets of data will have a smaller variance than does any estimator with less information of the process. The optimal estimator can be obtained by the maximum likelihood method because the maximum likelihood estimate has the minimum variance, at least for large samples, of all possible unbiased estimators of $\theta$. Now let us study the maximum likelihood estimate of $\theta$ under the most ideal situation, i.e., when all evolutionary events are observable.

**Lower bound of the variances when all evolutionary events are observable:** The proper likelihood function under this assumption is given by (8). The log-likelihood function is therefore

$$\begin{aligned} \log L &= c - (n-1)\log(4N_e) + \eta \log\mu \\ &\quad - \mu \sum_{m=2}^{n} m t_m - \frac{1}{4N_e} \sum_{m=2}^{n} m(m-1)t_m \\ &= c - (n-1)\log\theta + (\eta + n - 1)\log\mu \\ &\quad - \mu \sum_{m=2}^{n} m t_m - \frac{\mu}{\theta} \sum_{m=2}^{n} m(m-1)t_m \end{aligned} \tag{17}$$

where $c$ is not a function of $4N_e$ or $\mu$. From the log-likelihood function, we obtain the first order derivatives:

$$\frac{\partial \log L}{\partial \theta} = -\frac{n-1}{\theta} + \frac{\mu}{\theta^2} \sum_{m=2}^{n} m(m-1)t_m \tag{18}$$

$$\frac{\partial \log L}{\partial \mu} = \frac{\eta + n - 1}{\mu} - \sum_{m=2}^{n} m t_m - \frac{1}{\theta} \sum_{m=2}^{n} m(m-1)t_m. \tag{19}$$

Equating these two derivatives to zero and solving the equations for $\theta$ and $\mu$, we obtain the maximum likelihood estimates of $\theta$ and $\mu$ as

$$\hat{\mu} = \frac{\eta}{\sum_{m=2}^{n} m t_m} \tag{20}$$

$$\hat{\theta}_f = \frac{1}{n-1} \sum_{m=2}^{n} m(m-1)\hat{\mu} t_m \tag{21}$$

$$= \frac{\eta \sum_{m=2}^{n} m(m-1)t_m}{(n-1)\sum_{m=2}^{n} m t_m} \tag{22}$$

where the subscript $f$ of $\theta$ means that the estimation is made using the *full* information. To obtain the variance of the estimate, we need the second order derivatives. From (18) and (19), we have

$$\frac{\partial^2 \log L}{\partial \theta^2} = \frac{n-1}{\theta^2} - \frac{2\mu}{\theta^3} \sum_{m=2}^{n} m(m-1)t_m$$

$$\frac{\partial^2 \log L}{\partial \theta \partial \mu} = \frac{1}{\theta^2} \sum_{m=2}^{n} m(m-1)t_m$$

$$\frac{\partial^2 \log L}{\partial \mu^2} = -\frac{\eta + n - 1}{\mu^2}.$$

Fisher's information matrix is therefore given by

$$I(\theta, \mu) = -\begin{bmatrix} E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right) & E\left(\frac{\partial^2 \log L}{\partial \theta \partial \mu}\right) \\ E\left(\frac{\partial^2 \log L}{\partial \theta \partial \mu}\right) & E\left(\frac{\partial^2 \log L}{\partial \mu^2}\right) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{n-1}{\theta^2} & \frac{n-1}{\theta\mu} \\ \frac{n-1}{\theta\mu} & \frac{\theta a_n + n - 1}{\mu^2} \end{bmatrix}$$

The inverse of the information matrix can be easily obtained and from the inverse matrix, the asymptotic variance of $\hat{\theta}_f$ and $\hat{\mu}$ are found to be

$$\mathrm{Var}(\hat{\theta}_f) = \frac{\theta}{a_n}\left(1 + \frac{\theta a_n}{n-1}\right) \tag{23}$$

$$\mathrm{Var}(\hat{\mu}) = \frac{\mu}{a_n \theta}. \tag{24}$$

Note that if $\mu$ is known, then from (21), the variance of the estimate of $\theta$ will be

$$\mathrm{Var}(\hat{\theta}_f) = \frac{\mu}{n-1}\sum_{m=2}^{n} m^2(m-1)^2 \mathrm{Var}(t_m) = \frac{\theta^2}{n-1}. \tag{25}$$

Comparing this variance with (23), one can see that the first term, $\theta/a_n$, in (23) is due to the estimation of $\mu$.

**Lower bound of the variances when only the mutational events are observable:** Before discussing the implications of the results above, let us consider a less ideal situation in which the person, who can follow the course of evolution, observes (or scores) only the number of mutations in each branch of the genealogy between successive branching events. That is, he has information on $\{\eta_m, m = 2, \ldots, n\}$ but does not know the values of $\{t_m, m = 2, \ldots, n\}$. Let us see what his best estimator is.

The proper likelihood function is given by (10) and thus the log-likelihood function is

$$l = \log L = \sum_{m=2}^{n} \{\eta_m \log\beta_m - (\eta_m + 1)\log(\beta_m + 1)\}.$$

From this, it is easy to show

$$\frac{dl}{d\theta} = \sum_{m=2}^{n}\left(\frac{\eta_m}{\theta} - \frac{\eta_m + 1}{\theta + m - 1}\right). \tag{26}$$

Therefore, the maximum likelihood estimate of $\theta$ is the solution of the equation

$$\sum_{m=2}^{n} \frac{\eta_m + 1}{\theta + m - 1} = \frac{\eta}{\theta}. \tag{27}$$

This estimator will be denoted by $\hat{\theta}_m$. To calculate the variance of $\hat{\theta}_m$, note that

$$E\left(-\frac{\partial^2 l}{\partial \theta^2}\right) = E\left[\sum_{m=2}^{n}\left(\frac{\eta_m}{\theta^2} - \frac{\eta + 1}{(\theta + m - 1)^2}\right)\right]$$

$$= \sum_{m=2}^{n}\left(\frac{\theta}{\theta^2(m-1)} - \frac{\theta/(m-1) + 1}{(\theta + m - 1)^2}\right)$$

$$= \frac{a_n}{\theta} - \sum_{m=2}^{n} \frac{1}{(m-1)(\theta + m - 1)}.$$

The large sample variance of $\hat{\theta}_m$ is therefore

$$\mathrm{Var}(\hat{\theta}_m) = 1/E\left(-\frac{\partial^2 l}{\partial \theta^2}\right) = \frac{\theta}{a_n}\left(1 - \frac{\theta}{a_n}\sum_{k=2}^{n}\frac{1}{(k-1)(\theta + k - 1)}\right)^{-1} = \frac{\theta}{a_n}\left(1 + \frac{\alpha}{1-\alpha}\right) \tag{28}$$

where

$$\alpha = \frac{\theta}{a_n}\sum_{k=2}^{n}\frac{1}{(k-1)(\theta + k - 1)} = 1 - \frac{1}{a_n}\sum_{k=1}^{n-1}\frac{1}{\theta + k}.$$

**Implications:** We have established two lower bounds of the variances of estimators of $\theta$. As noted earlier, the variance of the best practical estimator can not be smaller than $\mathrm{Var}(\hat{\theta}_f)$ given by (23). In the real world we have far less information than that used to derive (23). Given a set of DNA sequences, one can at best reconstruct their genealogy without error, which includes the topology of the tree and the number of mutations in each branch. All other quantities such as $\{\eta_m, m = 2, \ldots, n\}$ and $\{t_m, m = 2, \ldots, n\}$ have to be estimated from the reconstructed genealogy. It should be emphasized that the information provided in the derivation of $\mathrm{Var}(\hat{\theta}_m)$ is actually more than that a perfectly reconstructed genealogy can provide. Therefore, we conclude that the variance of $\mathrm{Var}(\hat{\theta}_m)$ given by (28) is a lower bound of the variance of all possible estimators in practice. Note that when the number of genes in the sample is two ($n = 2$), all the estimators considered here, including TAJIMA's estimator, are identical and so are their variances.

Since $1/k > 1/(k+1)$, $k = 1, \ldots,$ Chebyshev's inequality ensures that

$$a_n^2 \leqslant (n-1)b_n.$$

It can further be shown that

$$\frac{\theta a_n}{n-1} \leqslant \frac{\alpha}{1-\alpha} \leqslant \frac{\theta b_n}{a_n}.$$

Therefore, from (14), (23) and (28), we have, as expected, that

$$\mathrm{Var}(\hat{\theta}_f) \leqslant \mathrm{Var}(\hat{\theta}_m) \leqslant \mathrm{Var}(\hat{\theta}_w)$$

We have considered here the estimates of $\theta$ for a

given locus. The parameter that is of most biological interest is the $\theta$ value per site, that is, $\theta^* = \theta/L$, where $L$ is the length of the sequences (number of nucleotides in a sequence). Corresponding to the three estimates $\hat{\theta}_f$, $\hat{\theta}_m$ and $\hat{\theta}_w$ of $\theta$, the estimates of $\theta^*$ are respectively $\hat{\theta}_f^* = \hat{\theta}_f/L$, $\hat{\theta}_m^* = \hat{\theta}_m/L$ and $\hat{\theta}_w^* = \hat{\theta}_w/L$. It follows from (14), (23) and (28) that

$$\mathrm{Var}(\hat{\theta}_w^*) = \frac{\theta^*}{a_n}\left(\frac{1}{L} + \frac{\theta^* b_n}{a^n}\right) \tag{29}$$

$$\mathrm{Var}(\hat{\theta}_f^*) = \frac{\theta^*}{a_n}\left(\frac{1}{L} + \frac{\theta^* a_n}{n-1}\right) \tag{30}$$

$$\mathrm{Var}(\hat{\theta}_m^*) = \frac{\theta^*}{a_n}\left(\frac{1}{L} + \frac{\alpha}{L(1-\alpha)}\right). \tag{31}$$

One can consider the ratio of the variances of two estimates of $\theta$ or $\theta^*$ (asymptotic relative efficiency) by letting the sample size $n$ or the sequence length $L$ or both approach infinity. For example, we have

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\hat{\theta}_m)}{\mathrm{Var}(\hat{\theta}_w)} = \lim_{n\to\infty} \frac{\mathrm{Var}(\hat{\theta}_f)}{\mathrm{Var}(\hat{\theta}_w)} = \lim_{n\to\infty} \frac{1 + \dfrac{\theta a_n}{n-1}}{1 + \dfrac{\theta b_n}{a_n}} = 1. \tag{32}$$

This means that for finite sequences WATTERSON's estimator $\hat{\theta}_w$ of $\theta$ has asymptotically the same variance as $\hat{\theta}_f$ and $\hat{\theta}_m$. In other words, WATTERSON's estimator of $\theta$ is asymptotically optimal and thus the single value of $\eta$ asymptotically contains all the information about the value of $\theta$.

Consider the same ratio but let the sequence length $L$ instead of the sample size $n$ approach infinity. If we assume that the $\theta$ has a finite limit $\theta_\infty$ when $L$ approach infinity (a common assumption for many study in population genetics). Then

$$\lim_{n\to\infty}\lim_{L\to\infty} \frac{\mathrm{Var}(\hat{\theta}_f^*)}{\mathrm{Var}(\hat{\theta}_w^*)} = \lim_{n\to\infty} \frac{1 + \dfrac{\theta_\infty a_n}{n-1}}{1 + \dfrac{\theta_\infty b_n}{a_n}} = 1. \tag{33}$$

However, if $\theta_\infty = \infty$ (that is, $\theta^*$ has a constant value), then the above limiting process becomes

$$\lim_{n\to\infty}\lim_{L\to\infty} \frac{\mathrm{Var}(\hat{\theta}_f^*)}{\mathrm{Var}(\hat{\theta}_w^*)} = \lim_{n\to\infty} \frac{\dfrac{a_n}{n-1}}{\dfrac{b_n}{a_n}} = 0.$$

The last result was obtained earlier by FELSENSTEIN (1992). On the other hand, if one let the sample size approach infinity first, then

$$\lim_{L\to\infty}\lim_{n\to\infty} \frac{\mathrm{Var}(\hat{\theta}_f^*)}{\mathrm{Var}(\hat{\theta}_w^*)} = 1.$$

This is true whether or not $\theta_\infty$ is finite. Since sequence length is finite and so is the value of $\theta$ in practice, the efficiency of WATTERSON's estimator of $\theta$ will increase when sample size is sufficiently large.

**Potential improvement:** We have established two lower bounds of the variances of all possible estimators of $\theta$ or $\theta^*$. However, because the variances of the two estimators $\hat{\theta}_f$ and $\hat{\theta}_m$, given by (23) and (28), respectively, are large sample variances, it is not apparent as to how different these variances can be when the sample size is small. To answer this question, computer simulations were conducted. Simulation results suggest that these large sample variance approximations are in fact very accurate for small samples as well. Two examples are given in Figure 2, one for $\theta = 3$ and the other for $\theta = 10$. In Figure 2, a and b, the variances calculated from simulated samples for WATTERSON's and TAJIMA's estimates of $\theta$ are also plotted for comparison, though their exact sample variances are known.

It should be emphasized that although it has been shown above that WATTERSON's estimates of $\theta$ or $\theta^*$ are asymptotically efficient for finite sequences, substantial improvement in the accuracy of estimate of $\theta$ or $\theta^*$ may be possible because the efficiency of WATTERSON's estimator approach one at a very slow rate. This can be seen from Figure 3. In Figure 3, we plot the efficiency of WATTERSON's estimator $\hat{\theta}_w$ relative to the estimator $\hat{\theta}_m$ for a number of values of $\theta$ against the number of sequences in the sample. As one can see, the common feature of these efficiency curves is that the efficiency decreases for each fixed value of $\theta$ relatively fast with the number of sequences in the sample to a stable level. The smaller the $\theta$ is, the fast this stable level is reached. With even the considerably large sample of 500 sequences, there is still no sign of increase in efficiency of WATTERSON's estimator. From this perspective, the asymptotic efficiency of WATTERSON's estimator is of little practical importance. The extent to which the improvement can be made depends on the value of $\theta$–the larger the value of $\theta$, the higher the potential improvement. When $\theta = 20$ and with 30 or more sequences, the variance of the best estimator of $\theta$ can be only half of the variance of WATTERSON's estimator of $\theta$. This will be indeed a substantial improvement if such an estimator can be found.

Now if only the total number of mutations in the genealogy of a sampled of $n$ genes is known, can one improve the estimation of $\theta$ over the estimator $\hat{\theta}_w$ without trying to recover the mutational events $\{\eta_m\}$ and coalescent events $\{t_m\}$? The answer is no! This is because in this situation the best one can do is replace

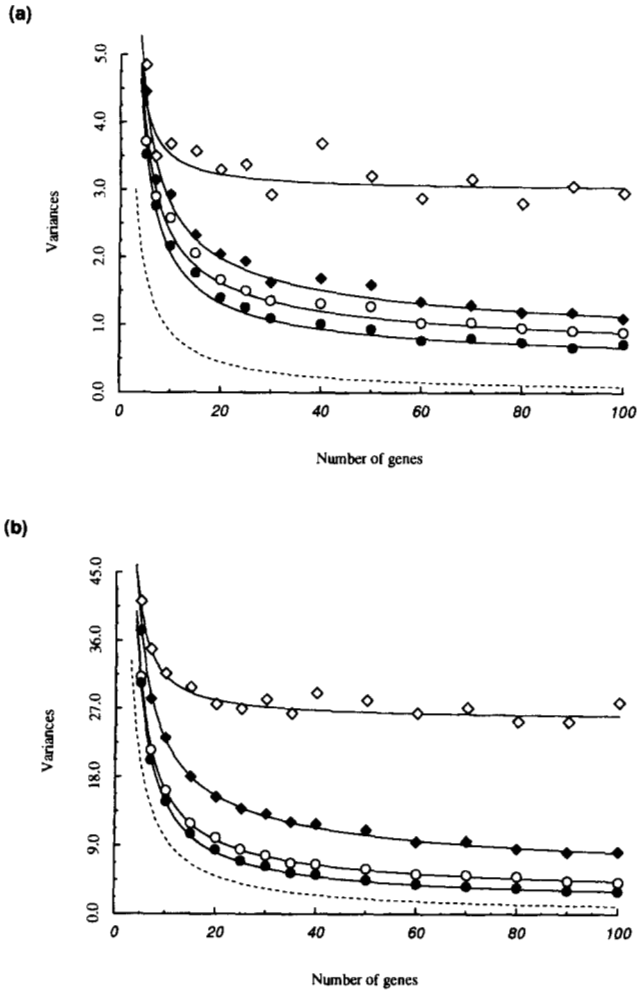


FIGURE 2.—Variances of estimates of $\theta$. The four solid curves from top to bottom are the theoretical variances of TAJIMA's estimate $\Pi_n$, WATTERSON's estimate $\hat{\theta}_w$ and the maximum likelihood estimates $\hat{\theta}_m$ and $\hat{\theta}_f$, respectively. The dashed curve is the variance given by (25). The points for each sample size (number of genes) are the variances from 5000 simulated samples. Notations: open diamond, TAJIMA's estimate; solid diamond, WATTERSON's estimate $\hat{\theta}_w$; open circle, the maximum likelihood estimate $\hat{\theta}_m$; and solid circle, the maximum likelihood estimate $\hat{\theta}_f$. (a) $\theta = 3.0$ and (b) $\theta = 10.0$.

the values of evolutionary events by their expected values conditional on the value of $\eta$. Note that

$$E(t_m) = \frac{4N_e}{m(m-1)}$$

$$E(\eta_m) = \beta_m$$

$$E(\eta_m|\eta) = E_{t_2,\ldots,t_n}E(\eta_m|\eta, t_2, \ldots, t_n)$$

$$= E_{t_2,\ldots,t_n}\left(\eta\frac{mt_m}{\sum_{m=2}^{n} mt_m}\right)$$

$$\approx \eta\frac{mE(t_m)}{\sum_{m=2}^{n} mE(t_m)} = \frac{\eta}{(m-1)a_n}.$$

Substituting $t_m$ in (22) by $E(t_m)$, we have

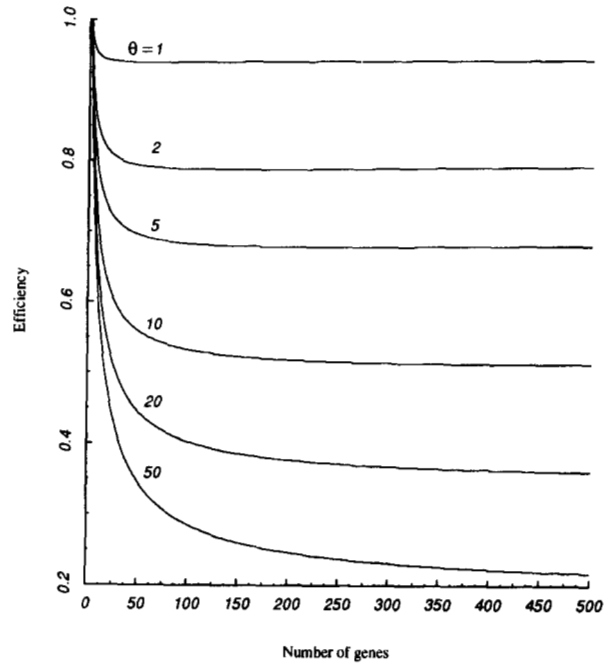


FIGURE 3.—Efficiency of WATTERSON's estimate $\hat{\theta}_w$ of $\theta$ relative to $\hat{\theta}_m$. The six curves from top to bottom correspond respectively to $\theta = 1$, 2, 5, 10, 20 and 50. The efficiency is calculated by $\mathrm{Var}(\hat{\theta}_m)/\mathrm{Var}(\hat{\theta}_w)$ where $\mathrm{Var}(\hat{\theta}_w)$ and $\mathrm{Var}(\hat{\theta}_m)$ are given by (14) and (28), respectively.

$$\hat{\theta}_f = \eta\frac{\sum_{m=2}^{n} 4N_e}{(n-1)\sum_{m=2}^{n}[1/(m-1)]4N_e} = \frac{\eta}{a_n}$$

and so $\hat{\theta}_f = \hat{\theta}_w$. Substituting $\eta_w$ in the left hand side of (27) with the unconditional expectation $E(\eta_m)$, we have

$$\sum_{m=2}^{n}\frac{\theta/(m-1)+1}{\theta+m-1} = a_n$$

and so $\hat{\theta}_m = \hat{\theta}_w$. Finally, if one substitutes $\eta_m$ in the left hand side of (27) with the conditional expectation $E(\eta_m|\eta)$, the resulting estimate $\hat{\theta}_m$ can be shown numerically to be almost identical to $\hat{\theta}_w$, whatever the value of $\theta$ is. Therefore, in the absence of knowledge about the mutational events $\{\eta_m\}$ and coalescent times $\{t_m\}$, $\hat{\theta}_w$ is the best estimate. Of course, a better estimate may be obtained if one has some knowledge about either $\{\eta_m\}$ or $\{t_m\}$.

## ESTIMATION OF λ

We now consider the situation in which the sites of the genes are classified into two classes. Again, it is assumed that there is no recombination. The simplest case of such a situation is to divide a sequence into two adjacent segments, likely with different lengths and mutation rates. Another case is that the sequence is a protein coding region. The need to consider two types of site is obvious because synonymous substitutions occur more frequently than nonsynonymous ones. Estimation of the ratio of two $\theta$'s serves two

purposes. First, the ratio of the two $\theta$'s may be of interest. Second, knowledge of the ratio can be used to improve the estimation of one of the two $\theta$'s. The latter will be illustrated later.

Without loss of generality, suppose that a sequence is divided into two segments with mutation rates $\mu_1$ and $\mu_2$, respectively. Let $\mu = \mu_1 + \mu_2$, $\lambda = \theta_1/\theta_2$ and $\theta = \theta_1 + \theta_2$. Then, we have

$$\theta_1 = \frac{\theta\lambda}{1+\lambda} \quad \text{and} \quad \theta_2 = \frac{\theta}{1+\lambda}. \tag{34}$$

Since we assume that there is no recombination, the two segments have the same genealogy. Let $n_m$, whose meaning is the same as before, be defined for the first segment and $\zeta_m$ be the corresponding variable for the second segment. Then the joint probability density function is

$$\prod_{m=2}^{n} p(\eta_m, \mu_1 m t_m) p(\zeta_m, \mu_2 m t_m) g\left(t_m, \frac{4N_e}{m(m-1)}\right)$$
$$= \left(\frac{\lambda}{1+\lambda}\right)^{\eta}\left(\frac{1}{1+\lambda}\right)^{\zeta} \prod_m \frac{\kappa_m!}{\eta_m!\zeta_m!} p(\kappa_m, \mu m t_m) g \cdot \left(t_m, \frac{4N_e}{m(m-1)}\right) \tag{35}$$

where $\kappa_m = \eta_m + \zeta_m$. The marginal distribution of mutational events is therefore the joint distribution.

$$\left(\frac{\lambda}{1+\lambda}\right)^{\eta}\left(\frac{1}{1+\lambda}\right)^{\zeta} \prod_m \frac{\kappa_m!}{\eta_m!\zeta_m!}\left(\frac{1}{\beta_m+1}\right)\left(\frac{\beta_m}{\beta_m+1}\right)^{\kappa_m}. \tag{36}$$

The log-likelihood of either of these two likelihood functions can be written as

$$l = \log L = c + \eta\log\lambda - (\eta+\zeta)\log(1+\lambda) \tag{37}$$

where $c$ is a constant independent of $\lambda$. Therefore the maximum likelihood estimate of $\lambda$ from either of the two density functions is easily found to be

$$\hat{\lambda} = \eta/\zeta. \tag{38}$$

The maximum likelihood estimate of $\theta$ can be obtained from the formulas in the previous section by combining the two segments into a single one. For example, $\hat{\theta}_w = (\eta + \zeta)/a_n$. Note that $\hat{\lambda}$ is not defined if $\zeta$ is equal to zero. Although the property of the maximum likelihood method ensures that $\hat{\lambda}$ is asymptotically unbiased, the speed of approaching the true value of $\lambda$ requires special attention. Using Taylor's expansion, one can show that

$$E\left(\frac{\eta}{\zeta}\right) = \frac{E(\eta)}{E(\zeta)} + \frac{E(\eta)}{E^3(\zeta)}\mathrm{Var}(\zeta) - \frac{1}{E^2(\zeta)}\mathrm{Cov}(\eta,\zeta) + O\left(\frac{1}{E^2(\zeta)}\right) \tag{39}$$

where the covariance of $\eta$ and $\zeta$ is

$$\begin{aligned}\mathrm{Cov}(\eta,\zeta) &= E(\eta\zeta) - E(\eta)E(\zeta)\\ &= E(E(\eta\zeta \mid t_2, \ldots, t_m)) - E(\eta)E(\zeta)\\ &= E\left(\mu_1\mu_2\left(\sum_m m t_m\right)^2\right) - E(\eta)E(\zeta)\\ &= \sum_{i,j}\frac{\theta_1}{i}\frac{\theta_2}{j} + \theta_1\theta_2 b_n - E(\eta)E(\zeta)\\ &= \theta_1\theta_2 b_n.\end{aligned} \tag{40}$$

Therefore,

$$E(\hat{\lambda}) = \lambda\left(1 + \frac{1}{E(\zeta)}\right) + O\left(\frac{1}{E^2(\zeta)}\right). \tag{41}$$

It follows that the bias of estimation approaches zero at the same speed as $1/E(\zeta) = 1/(\theta_2 a_n)$. Increasing the sample size does not help very much because $a_n$ increases slowly. A correction of the bias in estimation should be done in practice. From (41), an obvious correction is

$$\tilde{\lambda} = \hat{\lambda}/\left(1 + \frac{1}{E(\zeta)}\right) = \frac{\eta}{\zeta+1}. \tag{42}$$

Note that $\tilde{\lambda}$ is fully defined. Simulation (Table 1) shows that $\tilde{\lambda}$ gives quite reasonable results, particularly when $\theta$ is not too small. When $\theta$ and the number of genes in the sample are both small, $\tilde{\lambda}$ tends to underestimate $\lambda$. The following estimator seems to be slightly better than $\tilde{\lambda}$ (Table 1)

$$\tilde{\lambda}^* = \frac{\eta}{\zeta + 1 - \frac{1}{n}}. \tag{43}$$

To calculate the variance of $\tilde{\lambda}$, note that since

$$\lim_{n\to\infty}\tilde{\lambda} = \lim_{n\to\infty}\tilde{\lambda}^* = \lim_{n\to\infty}\hat{\lambda},$$

its asymptotic variance is equal to the asymptotic variance of $\hat{\lambda}$, which is $1/E(-\partial^2 l/\partial\lambda^2)$. From (37), it follows that

$$E\left(-\frac{\partial^2 l}{\partial\lambda^2}\right) = E\left(\frac{\eta}{\lambda^2} - \frac{\eta+\zeta}{(1+\lambda)^2}\right) = \frac{\theta_1}{\lambda^2} - \frac{\theta}{(1+\lambda)^2} = \frac{a_n\theta}{\lambda(1+\lambda)^2}.$$

Therefore, the large sample variance of $\tilde{\lambda}$ is

$$\mathrm{Var}(\tilde{\lambda}) = \frac{\lambda(1+\lambda)^2}{a_n\theta}. \tag{44}$$

This variance is in fact the lower bound of the variance

TABLE 1

**Estimates of λ from Equation 42**

| | λ = 0.2 | | | | λ = 0.5 | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $\phi = 1.0$ | $\phi = 2.0$ | $\phi = 4.0$ | $\phi = 10.0$ | $\phi = 1.0$ | $\phi = 2.0$ | $\phi = 4.0$ | $\phi = 10.0$ |
| 5 | 0.149 | 0.184 | 0.194 | 0.199 | 0.341 | 0.429 | 0.473 | 0.497 |
| | (0.170) | (0.201) | (0.204) | (0.203) | (0.396) | (0.477) | (0.502) | (0.509) |
| 10 | 0.177 | 0.196 | 0.200 | 0.199 | 0.409 | 0.477 | 0.494 | 0.497 |
| | (0.187) | (0.202) | (0.203) | (0.200) | (0.435) | (0.496) | (0.504) | (0.500) |
| 15 | 0.181 | 0.196 | 0.199 | 0.200 | 0.427 | 0.481 | 0.497 | 0.501 |
| | (0.187) | (0.200) | (0.201) | (0.201) | (0.444) | (0.492) | (0.502) | (0.503) |
| 30 | 0.193 | 0.198 | 0.198 | 0.199 | 0.453 | 0.484 | 0.500 | 0.499 |
| | (0.196) | (0.200) | (0.199) | (0.199) | (0.460) | (0.488) | (0.502) | (0.500) |
| 40 | 0.193 | 0.200 | 0.202 | 0.199 | 0.457 | 0.503 | 0.497 | 0.498 |
| | (0.195) | (0.201) | (0.203) | (0.199) | (0.463) | (0.506) | (0.498) | (0.498) |
| 50 | 0.192 | 0.200 | 0.199 | 0.201 | 0.467 | 0.494 | 0.499 | 0.498 |
| | (0.193) | (0.200) | (0.200) | (0.201) | (0.472) | (0.496) | (0.500) | (0.499) |
| 100 | 0.196 | 0.199 | 0.198 | 0.198 | 0.480 | 0.497 | 0.500 | 0.499 |
| | (0.197) | (0.199) | (0.198) | (0.198) | (0.482) | (0.498) | (0.501) | (0.499) |

Each entry in the table is the average from 10,000 simulations. The numbers in parenthesis are the estimates from Equation 43.

of all possible estimators of λ. Therefore, for small samples, the actual variance of $\tilde{\lambda}$ is expected to be larger than (44). In order to estimate the variance, the values of $\theta$ and λ in (44) have to be substituted with their estimates. In doing so, the estimated variance tends to be an overestimate, because it is the ratio of two non-negative random variables. Therefore, it is not straightforward to obtain a reliable estimate of the small sample variance of $\tilde{\lambda}$. Nevertheless, the variance may be approximated by the following modification

$$\frac{\tilde{\lambda}(1+\tilde{\lambda})^2}{a_n\hat{\theta}+c_n}$$

where $c_n$ is the correction factor. We found by simulation that $c_n$ should always be positive, implying that the inflation of estimation due to the ratio of two random variables does not entirely compensate the difference between small sample and asymptotic variances. Furthermore, we found that by choosing $c_n = a_n$, the estimated variance is reasonably accurate. We therefore suggest that for small samples ($n < 50$) the following formula be used to calculate the variance of $\tilde{\lambda}$.

$$\mathrm{Var}(\tilde{\lambda}) = \frac{\tilde{\lambda}(1+\tilde{\lambda})^2}{a_n\hat{\theta}+a_n}. \tag{45}$$

The variance of $\tilde{\lambda}^*$ should be computed by replacing $\tilde{\lambda}$ in (45) with $\tilde{\lambda}^*$

A computationally intensive method can be used to derive more accurate estimates of the variances for small samples. Since the variance of the estimate depends on only the sample size $n$ and the values of λ and $\theta$, once the estimates of the two parameters, λ and $\theta$, are obtained, one can simulate a large number of samples of size $n$ according to the distribution (6) and (7) using the estimated values of λ and $\theta$. Let $\lambda_i$ be the estimate of λ based on the $i$-th simulated sample and $S$ be the number of samples simulated. Then the variance of the estimate of λ is estimated by the sample variance of $\lambda_i$'s. That is,

$$\mathrm{Var}(\tilde{\lambda}) = \frac{1}{S-1}\sum_{i=1}^{S}(\lambda_i - \bar{\lambda})^2.$$

This is the so-called parametric bootstrap variance estimate.

Let us now see how a knowledge of λ can help improve the estimate of the $\theta$ in a segment by incorporating information from the other segment. Suppose we are interested in the value of $\theta_1$ but the value of λ is known or can be assumed to be approximately the ratio of the two segment lengths. Two estimates are possible. One is to estimate $\theta_1$ as if we do not have any other information. The other one is first to estimate $\theta$ and then use $\theta_1 = \theta\lambda/(1+\lambda)$ to obtain the estimate of $\theta_1$. The ratio of the variance of the latter estimate to that of the former is

$$\frac{\lambda}{1+\lambda}\left(\frac{1+\theta b_n}{1+\theta_1 b_n}\right) = \frac{\dfrac{\theta_1 b_n}{1+\theta_1 b_n}+\lambda}{1+\lambda} < 1.$$

We can see that using full information can always improve the accuracy of estimation. The gain is particularly substantial when $\theta_1 \ll 1$ and $\lambda \ll 1$. For example, when $\theta_1 = 0.1$, $\lambda = 0.3$ and $n = 5$, the ratio of variances is 0.33. That is, the variance of the estimate using all the information is only 1/3 of the variance which does not use the knowledge about λ. In practice, of course, the value of λ will not be known exactly, but the estimate of λ can be extended with little difficulty to more complex situations. Since λ does not depend on population size and is insensitive

TABLE 2

$\phi_q$ defined by Equation 46

| $n$ | $q = 0.05$ | $q = 10^{-2}$ | $q = 10^{-3}$ | $q = 10^{-4}$ | $q = 10^{-5}$ |
|---|---|---|---|---|---|
| 2 | 40.000 | 90.512 | 310.127 | 990.501 | 3150.728 |
| 3 | 10.837 | 30.713 | 90.033 | 200.560 | 450.423 |
| 4 | 0.915 | 10.861 | 40.235 | 80.562 | 160.318 |
| 5 | 0.399 | 0.922 | 20.235 | 40.469 | 80.100 |
| 6 | 0.126 | 0.381 | 10.140 | 20.461 | 40.529 |
| 7 | $0.261 \times 10^{-1}$ | 0.107 | 0.486 | 10.285 | 20.573 |
| 8 | $0.393 \times 10^{-2}$ | $0.189 \times 10^{-1}$ | 0.140 | 0.560 | 10.366 |
| 9 | $0.495 \times 10^{-3}$ | $0.246 \times 10^{-2}$ | $0.233 \times 10^{-1}$ | 0.162 | 0.600 |
| 10 | $0.551 \times 10^{-4}$ | $0.275 \times 10^{-3}$ | $0.273 \times 10^{-2}$ | $0.256 \times 10^{-1}$ | 0.173 |

to change in mutation rate as long as $\mu_1$ and $\mu_2$ are changed at about the same rate, it can be expected to be relatively stable across species. Therefore, the estimation of $\lambda$ can be done with high precision when multiple species data are available. Improving the accuracy of estimating $\lambda$ is thus an efficient way to improve the accuracy of the estimation of $\theta$'s.

## DISCUSSION

The estimation of $\theta$ in the special case that the observed $K$ is 0 is worth of further discussion. In this case, both WATTERSON's estimate and TAJIMA's estimate of $\theta$ are zero and so are the variances. This is not reasonable. It is more informative in this situation to give an interval estimate of $\theta$. WATTERSON (1975) found that

$$Pr(K = 0|\theta) = \frac{1}{(\theta+1)\ldots(\theta+n-1)}.$$

From this equation, one can find the minimum value $\theta_q$ of $\theta$ such that

$$Pr(K = 0|\theta_q) \leq q. \tag{46}$$

The interval estimate of $\theta$ is then $(0, \theta_q)$. Table 2 gives the $\theta_q$'s for $q = 0.05$, $10^{-2}$, $10^{-3}$, $10^{-4}$ and $10^{-5}$. If $q = 0.05$, for example, then $\theta_q = 0.399$ for $n = 5$ and $\theta_q$ - 0.021 for $n = 10$.

J. FELSENSTEIN and R. HUDSON (personal communication) suggested to consider $p = \theta_1/(\theta_1 + \theta_2)$ instead of $\lambda = \theta_1/\theta_2$. The mathematics dealing with $p$ is rather simple because the random variable $\eta$ given the value of $\eta + \zeta$ follows a binomial distribution with parameter $\theta_1/(\theta_1 + \theta_2)$. Therefore $\hat{p} = \eta/(\eta + \zeta)$ is an unbiased estimator of $p$ and $\mathrm{Var}(\hat{p}) = p(1-p)/(\eta + \zeta)$.

Recall that $\theta_1$ and $\theta_2$ are defined for the whole segments. These two segments may have different number of sites. Therefore, it would be more appropriate to compare the mutation rates of the two segments on the per site basis. To achieve this, only a minor change is required. Let the ratio of the number of sites of the first segment to that of the second segment be $\rho$. Then the $\lambda$ per site is simply

$$\lambda_{\text{per site}} = \lambda/\rho.$$

In the estimation of $\lambda$, we have assumed up to now that the two segments in consideration are completely linked. If this is not the case, the genealogy of the two segments may be different. For the extreme case in which the two segments are unlinked (i.e., the recombination rate equals 0.5), the two genealogies are independent. In general, let the coalescent times for the segments be $\{t_m\}$ and $\{t'_m\}$, respectively, and their joint density function be $h(t_m; t'_m, m = 2, \ldots, n)$. Then the overall joint density function is

$$\left(\prod_{m=2}^{n} p(\eta_m; \mu_1 t_m) p(\zeta; \mu_2 t'_m)\right) h(t_m; t'_m, m = 1, \ldots, n-1)$$

which can be rewritten as

$$\left(\frac{1}{1+\lambda}\right)^{\eta}\left(\frac{\lambda}{1+\lambda}\right)^{\zeta} \left[\prod_{m=2}^{n} \exp(-\mu_1 m t_m)(-\mu_2 m t'_m)\right] g(\theta, \kappa's, t'_m s).$$

From this, it follows that the maximum likelihood estimate is the solution of the equation:

$$\frac{\partial \log L}{\partial \lambda} = \frac{\zeta}{\lambda} - \frac{\eta+\zeta}{1+\lambda} + \frac{1}{(1+\lambda)^2}\sum_m m(t'_m - t_m) = 0.$$

In the absence of information about the values of $t_m$ and $t'_m$ ($m = 2, \ldots, n$), the best one can do is to replace them by their expected values. This will then lead to the maximum likelihood estimate $\hat{\lambda} = \eta/\zeta$. Again bias correction should be done. It is easy to see that the large sample variance of the maximum likelihood estimate of $\lambda$ is also given by (44), provided that $E(t_m) = E(t'_m)$. This is true if the sequences in the sample are randomly chosen. The case where the two segments are completely linked corresponds to $t_m = t'_m, m = 2, \ldots, n$.

Although we derived lower bounds of the variances for the estimator of $\theta$, it is not clear so far on how

close the variance of a practical estimator to the lower bound $\mathrm{Var}(\hat{\theta}_m)$ can be. However, we believe that an accurate reconstruction of gene genealogy will eventually lead to an estimator of $\theta$ with a variance close to the lower bound, because the number of mutations on each branch and the branching order of the nodes should provide almost as much information as the values of $\{\eta_m, m = 1, \ldots\}$. Such an estimator may perhaps be developed along the line of work by STROBECK (1983), ETHIER and GRIFFITH (1987) and GRIFFITH (1989). In particular, if the probability distribution of samples computed in GRIFFITH (1989) is adapted for estimating $\theta$, we expect that the variance of the estimate be close to the lower bound $\mathrm{Var}(\hat{\theta}_m)$.

We thank J. FELSENSTEIN for giving us a copy of his paper before publication. We also thank J. FELSENSTEIN and R. HUDSON for suggestions. This study was supported by National Institutes of Health grants.

## LITERATURE CITED

ETHIER, S. N., and R. C. GRIFFITHS, 1987 The infinitely-many-sites model as a measure-valued diffusion. Ann. Prob. **15:** 515–545.

FELSENSTEIN, J. 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregation sites as compared to phylogenetic estimates. Genet. Res. **56:** 139–147.

GRIFFITHS, R. C., 1989 Genealogical tree probabilities in the infinitely-many-site model. J. Math. Biol. **27:** 667–680.

HUDSON, R., 1982 Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

KINGMAN, J., 1982 On the genealogy of large populations. J. Appl. Prob. **19A:** 27–43.

STROBECK, C., 1983 Estimation of the neutral mutation rate in a finite population from DNA sequence data. Theor. Popul. Biol. **24:** 160–172.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

WATTERSON, G., 1975 On the number of segregation sites. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: R. R. HUDSON